(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
4 April 2002 (04.04.2002)

PCT

(10) International Publication Number
WO 02/27327 A2

(51) International Patent Classification[7]: G01N 33/68

(21) International Application Number: PCT/GB01/03693

(22) International Filing Date: 17 August 2001 (17.08.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0020357.0 17 August 2000 (17.08.2000) GB
60/247,995 14 November 2000 (14.11.2000) US

(71) Applicant *(for all designated States except US)*: **SENSE PROTEOMIC LIMITED** [GB/GB]; The Babraham Bioincubator, Babraham Hall, Babraham, Cambridgeshire CB2 4AT (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **BLACKBURN, Jonathan, Michael** [GB/GB]; 36 Woodlark Road, Cambridge CB3 0HS (GB). **MULDER, Michelle, Anne** [ZA/GB]; 83 Beche Road, Cambridge CB5 8HU (GB). **SAMADDAR, Mitali** [IN/GB]; 40 Chatsworth Avenue, Cambridge CB4 3LT (GB). **KOZLOWSKI, Roland** [GB/GB]; Sense Proteomic Limited, The Babraham Bioincubator, Babraham Hall, Babraham, Cambridge CB2 4AT (GB).

(74) Agents: **CHAPMAN, P., W.** et al.; Kilburn & Strode, 20 Red Lion Street, London WC1R 4PJ (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/27327 A2

(54) Title: METHOD

(57) Abstract: The present invention relates to novel methods of producing proteins in which one or more domains are full length and correctly folded and which are each tagged at either the N- or C-terminus with one or more marker moieties and arrays containing such proteins, as well as the use of such proteins in arrays for rapid screening.

METHOD

The present invention relates to novel methods of producing proteins in which one or more domains are full length and correctly folded and which are each tagged at either the N- or C-terminus with one or more marker moieties and arrays containing such proteins, as well as the use of such arrays in rapid screening.

The genome mapping projects are revolutionising the therapeutic target discovery process and with it the drug discovery process. As new therapeutic targets are identified, high throughput screening of existing and combinatorial chemical libraries will suggest many potential lead compounds which are active against these targets. It will clearly be uneconomic to pursue all lead compounds through even early phase clinical trials; currently however no rapid method exists for evaluating such lead compounds in terms of their likely activity profiles against all proteins in an organism. If available, such a method would allow the potential toxicology profiles of all the lead compounds to be assessed at an early stage and this information would significantly enhance the process of deciding which lead compounds to pursue and which to set aside.

There is a complementary need in the pharmaceutical industry to identify all the targets of existing drugs (either in the market already or still in development) and hence to define their mechanism of action. The availability of such information will greatly facilitate the process of gaining regulatory approval for new drugs since it is increasingly clear that the regulatory bodies now regard a knowledge of the mechanism of action to be of paramount importance. In addition, this type of information would enable the design of improved second generation drugs. This follows because the majority of drugs have at least minor side effects, which probably result from binding of the drug or a metabolite thereof to undesirable targets; all of these target proteins need to be identified in order to define the criteria necessary for design of improved drugs. Currently however no simple method exists to generate this information and a number of potential multi-million dollar drugs fall by the wayside

simply for lack of knowledge of the target of action.

Protein-protein interactions are being increasingly recognised as being of critical importance in governing cellular responses to both internal and external stresses. Specific protein-protein interactions therefore represent potential targets for drug-mediated intervention in infections and other disease states. Currently the yeast two-hybrid assay is the only reliable method for assessing protein-protein interactions but *in vivo* assays of this type will not be readily compatible even in a non-high throughput format with the identification of specific agonists or antagonists of protein-protein interactions. Functional proteome expression arrays, or "proteome chips", will enable the specificity of protein-protein interactions and the specificity of any drug-mediated effect to be determined in an *in vitro* format. They will therefore have enormous potential because they will simply revolutionise this area of research.

One way in which functional proteome arrays could be generated is to individually clone, express, purify and immobilise all proteins expressed in the specific proteome. Here though, an important initial consideration concerns the absolute size of the genome of interest together with considerations about the availability of sequence data for the entire genome.

By way of illustration of these points, a typical bacterial genome is ~5Mbp and a small number have now been completely sequenced (for example *Helicobacter pylori*, *Escherichia coli*, and *Mycobacterium tuberculosis*); fungal genomes are typically ~40Mbp, mammalian genomes at ~3Gbp and plant genomes at ~10Gbp. Current estimates are that the human genome sequence will be finished around 2003, although how much of this information will be in the public domain is very much open to question. Clearly it will be completely impractical to expect that the genomes of anything other than representative model organisms will become available in a realistic time frame, yet from the perspective of functional proteomics, model organisms are of only limited value. So, whilst in principle within the next four years

it may be possible to design and synthesise primers to clone each of the ~100,000 genes in the human genome from cDNA libraries, in practice this will be both enormously expensive (the cost of primers alone would run in to several millions of dollars) and a hugely laborious process, even if the necessary sequence data is available.

But what about those pharmaceutically relevant organisms for which the complete sequence data will not be available? These cannot be simply ignored by functional proteomics so what are the alternatives? Expression cDNA libraries could in principle be used together with non-specific immobilisation to create an array of proteins, but this technology is significantly limited by the fact that non-specific immobilisation is usually associated with loss of function because the fold of the protein is disrupted. In addition, all host cell proteins will also be immobilised which will at best markedly reduce signal-to-noise ratios and at worst result in obfuscation of positive results. The ability to create a functional proteome array in which individual proteins are specifically immobilised and purified *via* a common motif or tag without affecting function and without requiring knowledge of the entire genome sequence would therefore represent a huge advance in the field of functional proteomics.

The Inventors have now developed a novel approach which solves the problems described above by providing methodology which allows each protein in a proteome to be tagged with a common marker at a defined position within the protein without requiring any prior knowledge of the DNA sequence of the corresponding genes. This 'tag' can then be used to impart a commonality and specificity to downstream immobilisation and purification procedures, which in turn enables the creation of spatially defined arrays in which many thousands of proteins from a given proteome are displayed.

An important consideration here relates to the precise positioning of the 'tag'. If the tag is inserted in-frame in to any gene at an undefined, random position, the likelihood

is that the resultant tagged protein will be truncated in an undefined manner and in the majority of cases correct folding, and hence function, will be destroyed. It is often found that full-length proteins have short polypeptide extensions at either (or both) the N- and C-termini which can be truncated without affecting folding or function. However, if the truncations remove any N- or C-terminal extensions and cross a domain boundary, folding and function of the protein are usually then compromised. The methodology described here allows the tag to be inserted in the correct reading frame either precisely at the N- or C-terminus of each protein, or within a region close to either terminus which is unimportant in the folding and function of the protein, such that the individual tagged proteins fold correctly and hence retain function when specifically immobilised in the array. In the case of multidomain proteins where individual domains have discrete functions, the methodology described here also allows insertion of the tag within the overall coding sequence but outside specific domain boundaries such that the individual tagged domains fold correctly and hence retain function when specifically immobilised in the array.

Since each protein in the array will be fully functional, the arrays can then be screened directly to identify the targets of drugs and other biologically relevant molecules. The spatial definition of the arrays will allow the phenotype of each protein to be related directly to its genotype to allow the identification of 'hits'.

Thus, in a first aspect, the present invention provides a method producing one or more proteins in which one or more domains are full length and correctly folded and which are each tagged at either the N- or C-terminus with one or more marker moieties, said method comprising:

(a) providing one or more DNA molecules having an open reading frame encoding said proteins together with 5' and/or 3' untranslated regions;

(b) amplifying said DNA molecules under conditions that statistically incorporate α-S-dNTPs as well as dNTPs into the daughter DNA molecules;

(c) specifically protecting the 5' or 3' end of said DNA molecules from nuclease digestion;

(d) treating said DNA molecules first with a 5' to 3'- or 3' to 5'-nuclease to generate a set of nested deletions followed by treating with a single-strand nuclease under conditions that allow removal of said 5' or 3' untranslated regions including the start or stop codons of said open reading frame;

(e) cloning the fragments generated by step (d) into an expression vector containing a coding sequence for one or more 5' or 3' marker moieties;

(f) expressing said encoded proteins.

Preferably the amplification of the DNA molecule or molecules statistically incorporates a single α-S-dNTP, more preferably eiither α-S-dTTP or α-S-dATP.

The marker moiety can be either a peptide sequence, eg a hexa-histidine tag, an antibody epitope or a biotin mimic, or indeed a complete protein, or protein domain, eg the maltose binding protein domain. The marker moiety itself can be post-translationally modified, eg by addition of a biotin or lipid molecule. In a preferred embodiment, the marker moiety would also allow purification of "tagged" proteins.

Thus, the methods of the present invention allow the specific modification, in one pot, of every member of a cDNA library in a manner which does not rely on any knowledge of the sequence of individual genes. Instead, it relies on non-processive truncation of each cDNA by a nuclease such that either the 5'- or the 3'- untranslated region of each cDNA is removed. Additional known DNA sequence encoding a known marker moiety is then appended to the resultant set of nested deletions of each cDNA. If the marker moiety is in the same reading frame as the individual cDNA and is not preceded by any in-frame stop codon, each resultant genetically modified cDNA produced according to the methods of the present invention will thus encode an individual protein which now has a common moiety, eg. a polypeptide "tag" fused to

either its N- or C-terminus. A screen for correctly folded, tagged proteins then allows all truncations which cross a domain boundary and affect the folding (and hence function) of the individual protein and all out-of-frame fusions to the tag to be discarded.

Since every member of a cDNA library will be modified in the same manner, the net result will be that every protein encoded by the cDNA library will now be tagged with a common moiety at either their N- or C-termini.

In general, the proteins expressed from the cDNA library will be "tagged" and can be readily identified and isolated. Once purified they can be attached to microarrays, for example. Attachment can be effected by means of the tag itself, or alternatively, by means of another moiety which is first attached to the proteins.

Arrays formed by the methods described herein form a second aspect of the invention. Such arrays comprise the "tagged" protein expression library, immobilised, usually on a solid support. The skilled person will understand that a range of possible solid supports are in common usage in the area of arrays and any of these "substrates" can be utilised in the production of arrays of the present invention.

As discussed herein the term "protein array" relates to a spatially defined arrangement of one or more protein moieties in a pattern on a surface. Preferably the protein moieties will be attached to the surface either directly or indirectly. The attachment can be non-specific (e.g. by physical absorption onto the surface or by formation of a non-specific covalent interaction). In a preferred embodiment the protein moieties will be attached to the surface through the common marker moiety linked to each protein using the methods described herein.

In another embodiment, the protein moieties may be incorporated into a vesicle or liposome which is tethered to the surface.

Thus, for example, each position in the pattern may contain one or more copies of:

a) a sample of a single protein type (in the form of a monomer, dimer, trimer, tetramer or higher multimer);

b) a sample of a single protein type bound to an interacting molecule (e.g. DNA, antibody, other protein);

c) a sample of a single protein type bound to a synthetic molecule (e.g. peptide, chemical compound); or

d) mixtures of between 2 and 100 different tagged protein moieties at each position in the pattern of the array.

The surface which supports the array may be coated/derivatised by chemical treatment, for instance. Examples of suitable surfaces include glass slides, polypropylene or polystyrene, silica, gold or metal support or membranes made of, for example, nitrocellulose, PVDF, nylon or phosphocellulose.

As discussed herein, the methods of the present invention allow tagging of all proteins in a given proteome specifically at either the N- or C-terminus. Whilst some proteins may not tolerate N-terminal extensions and others might not tolerate C-terminal extensions, it is likely that the vast majority of proteins will tolerate one or other such extensions. Existing library cloning methods, however, simply cannot address this problem since they clone genes either as full-length, unmodified cDNAs or as random and almost inevitably truncated fusions to some protein partner. Compared to the latter, the present methods allow the position of the tag to be targeted to the sequences at or close to the N- or C-terminal residues of the cDNA products such that fusion to eg. a desired peptide partner does not affect folding or function of the cDNA product. Compared to the former, the method of immobilising proteins in an array as described

herein is through specific rather than non-specific interactions, and these specific interactions are a function of the tag added to the termini of each cDNA. Additionally, the methods described herein can be used to screen purified, immobilised proteins which have been expressed in non-bacterial host organisms to aid maintenance of function through correct folding and post-translational modification, whereas existing methods such as phage display or λ-cDNA expression libraries are restricted to bacterial hosts in which the majority of eukaryotic proteins are found to be synthesised in a non-functional form, either due to mis-folding or incorrect post-translational modification.

The methods of the present invention have a wide range of potential *in vitro* applications, which can be broadly divided into three main areas. These are the study of protein-ligand interactions, the study of protein-protein interactions, and the study of protein-DNA interactions.

Protein-Ligand Interactions

The methods described herein will allow the rapid profiling of the interactions between a given new chemical entity and all proteins in a given proteome. This can be achieved simply through probing the appropriate proteome array with the NCE at varying stringencies in what might be considered a reverse high throughput screen. The readout from such a screen will be directly useful in many situations, some of which are described below.

High throughput screening programs in which libraries of compounds are tested against cells or whole organisms often identifies leads, which give rise to a phenotypic change without the target being known prior to screening. Subsequent identification of the primary target can, however, be a very laborious process. The methods of the present invention can be applied directly to this type of problem since it will be possible to create a functional proteome array for the species concerned and then screen this array with the lead compound to identify which proteins within the

proteome it is targeting. This massively parallel approach to identifying protein-ligand interactions will greatly speed up and simplify the determination of primary targets of NCEs, and will also allow identification of weaker secondary interactions which may also be important. In addition, the methods can be applied directly to the question of species cross-reactivity, allowing a potential antifungal compound, for example, to be quickly assessed in terms of its interactions with, for example, all proteins in a human proteome; this type of information is likely to prove very useful in any subsequent optimisation of lead compounds.

High throughput screening methods now allow the rapid identification of small molecules which bind to a given protein which has itself previously been identified as a potential therapeutic target. However, these methods do not address the question of how selective any given interaction might be yet this knowledge is potentially crucial in deciding whether to pursue a given lead compound or not; perceived wisdom would argue that compounds which target single poteins are likely to show fewer side effects than those which also hit a large number of related or unrelated proteins.

There are a number of examples of compounds which have progressed successfully through third phase clinical trials yet have failed to win regulatory approval because their primary mechanism of action is not known. The antidepressant drugs mianserin and trazadone and the Pfizer anti-arthritic drug tenidap are examples here, each representing hundreds of millions of dollars investment for no return. The methods described herein can potentially be applied to the resurrection of such failed drugs since if the primary targets of such drugs can be discovered and subsequently verified in terms of mechanism of action, the vastly expensive clinical trial data is already in place for regulatory approval.

All existing drugs have side effects, to a greater or lesser extent, an example here being the otherwise attractive anti-schizophrenia drug clozapine. If the molecular origin of such side effects could be determined, this would greatly facilitate the design of future generation drugs with optimised primary effects combined with minimised

side effects. Again the presently described methods can be applied directly to such problems since in creating a profile of the interactions between a compound and all proteins in a proteome, aberrant secondary interactions will be identified and these can subsequently be assessed in terms of whether they are linked to known side effects.

The methods of the present invention can also be used to identify families of proteins, such as serine proteases, through screening proteome arrays with generic inhibitors. This would then allow the subsequent development of biochips displaying, for example, all human serine proteases or, alternately, all kinases or all p450 enzymes for more focused screening of lead compounds. A p450 biochip, for example, would have utility in assessing whether a given lead compound is likely to be metabolised or not, since p450-mediated hydroxylation is often the first step in this process and is thought to be one of the primary sources of patient-to-patient variability in drug response; indeed one of the goals of drug design now is to generate compounds which are not metabolised in the first place and here again a p450 chip would have significant potential utility.

Protein-protein interactions

Protein-protein interactions and multiprotein complexes are of critical importance in cellular biology. Signalling pathways, for example, are commonly initiated by an interaction between a cell surface receptor and an external ligand, and this is followed by a cascade of protein- protein interactions which ultimately result in the activation of a specific gene. Individual protein-protein interactions might be dependent on the presence of a specific ligand or alternatively might be blocked by a specific ligand, whilst some multiprotein complexes will only form in a ligand-dependent manner.

Thousands of new protein-protein interactions have been identified using two-hybrid technologies. The methods described herein overcome the limitations of such methods and can be used to screen proteome arrays with individual labelled proteins to identify

not only interacting partners but also the relative strengths of individual interactions. The methods can also be applied to the identification of the components of multiprotein complexes, even where their assembly is ligand dependent.

An example of the use of the methods in this way in defining novel protein-protein interactions would be the identification of the signalling partners of the cytosolic domain of a particular cell surface receptor which has been implicated in a disease state; identification of such signalling partners would be directly relevant from a pharmaceutical perspective since such protein-protein interactions might immediately represent possible therapeutic targets.

Protein-DNA Interactions

It has been estimated that roughly 10% of all genes in the human genome encode transcription factors yet only a small percentage of these are at present identified. The binding of specific transcription factors to DNA enhancer elements, often in response to external stimuli, is a prerequisite for the formation of enhanceosome complexes which then switch on gene expression. There are various points at which gene expression can in principle be affected by drug administration: a drug might block the binding of a protein or small molecule to a cell surface receptor and hence block the signalling cascade at the beginning; a drug might block a protein-protein interaction or inhibit an enzymatic activity within the signalling cascade; or altematively, a drug might block formation of specific protein-DNA or protein-protein interactions within the enhanceosome complex. As an example here, the transcription factor NF-κB is involved in cellular processes as diverse as immune and inflammation responses, limb development, septic shock, asthma, and HIV propeptide production. The majority of the intracellular signalling cascades in NF-κB activation are common to all these process so do not represent viable targets for intervention. The differences between the responses therefore lie in either the original ligand-receptor interaction or in the formation of specific enhanceosome complexes. NF-κB is known to bind to at least 14 different enhancer elements and the enhanceosome complexes therefore represent

potential therapeutic targets. However, delineation of an individual enhanceosome complex requires knowledge of both the number of individual DNA-binding proteins involved and also their protein-protein interactions with each other. The present methods can be used to directly address both these questions. A proteome array can be screened with specific DNA probes to identify novel DNA binding proteins, Alternatively, the proteome array can be screened with the transactivation domain of a given transcription factor to identity other proteins with which it interacts. Cross correlation of such screens should allow identification of new components of specific enhanceosome complexes

The protein arrays generated by the methods of the present invention will also allow the selection of molecules which recognise each protein displayed in the arrays. In a preferred embodiment, the selected molecules will be antibodies or antibody-like proteins and will be displayed on phage or on ribosomes or will be covalently linked to the encoding mRNA.

Thus, a phage displayed antibody library can be applied to each immobilised protein in the array and non-binding antibodies removed by washing. The selected phage can then be recovered and used to infect bacteria according to normal procedures. The phage-infected bacteria can then produce either phage particles displaying the selected antibodies for further rounds of selection, or they can produce soluble antibody fragments for direct use. The terms 'antibody' or 'antibody fragments' here refer to single chain Fvs, FAB fragments, individual light or heavy chain fragments, derived from mouse, human, camel or other organisms.

In a preferred embodiment, the protein array will be in microwell format such that after the selection step, the phage particles can be recovered by addition of appropriate bacterial cells to each well where they will become infected by the selected phage particles. Growth media can then be added to each well and the infected bacteria allowed to grow and express the antibody fragments, whilst maintaining the physical

separation of the antibody fragments selected to each immobilised protein in the array. If so desired, new phage particles produced by the infected bacteria can be used in subsequent rounds of selection. Such procedures are now routine for selecting polyclonal or monoclonal antibody fragments to a single purified and immobilised protein. In effect then the original protein arrays here will allow the generation of polyclonal or monoclonal antibody fragments to thousands of correctly folded proteins in a massively parallel manner whilst otherwise using standard in vitro antibody selection methods.

The selected, solubly expressed antibody fragments from each well of the original array can themselves be immobilised in to a new spatially defined array such that the antibody fragments in each position of the new array were selected against the proteins immobilised in a single, defined position in the original array. The antibody arrays so-generated will contain at each position either polyclonal or monoclonal antibody fragments, depending on the number of rounds of selection carried out prior to immobilisation of the soluble antibody fragments.

Such antibody arrays will have a number of potential uses including capture of individual proteins from a crude cell or tissue lysate for differential expression monitoring of the relevant proteome. Alternatively, the antibody-captured proteins might be screened directly for ligand-binding function. In general, any one monoclonal antibody might bind to the target protein so as to block its function, but another monoclonal antibody might bind but not block function. In a massively parallel approach, it is clearly impractical to assess all monoclonal antibodies to all proteins in a proteome individually for their ability to bind but not affect function. A polyclonal set of antibodies to all proteins in a proteome however is likely to contain individual antibodies which have the desired ability to bind but not affect function and will, in addition, contain individual antibodies which recognise all post-translational modifications of a given protein. Thus in general, polyclonal rather than monoclonal

antibody arrays generated as described will likely be advantageous for screening captured proteins directly for function.

Compared to the original protein arrays, the antibody arrays created by the methods described here will have the advantage that all proteins immobilised on the array will be stable under similar conditions. The proteins captured from the crude cell or tissue lysate will not be recombinant but will have been naturally expressed. Moreover, the captured proteins can be screened for function or ligand binding *etc* directly after capture from the crude cell or tissue lysate, which should aid maintenance of function.

Thus, in further aspects, the present invention provides:

(i) a method of screening one or more compounds for biological activity which comprises the step of bringing said one or more compounds into contact with a protein array as defined herein and measuring binding of the one or more compounds to the proteins in the array;

(ii) a method of screening one or more proteins for specific protein-protein interactions which comprises the step of bringing said one or more proteins, eg a cell surface receptor, into contact with an array as defined herein, and measuring binding of the one or more specific proteins with the proteins of the array;

(iii) a method of screening one or more proteins for specific protein-nucleic acid interactions which comprises the step of bringing said one or more nucleic acid probes into contact with an array as defined herein and measuring binding of the probes to the proteins in the array.

(iv) the use of an array as defined herein in the rapid screening of a compound, protein or nucleic acid;

(v) the use of an array as defined herein in screening for molecules which recognise each protein in the array, wherein the molecules are preferably antibodies;

(vi) a method of generating an antibody array which comprises bringing a protein array, as defined herein, into contact with an antibody library, such that one or more proteins in the protein array bind to at least one antibody in the antibody library, removing any unbound antibodies and immobilisation of those antibodies bound to proteins in the protein array; and

(vii) a method for the screening of protein function or abundance which comprises the step of bringing an antibody array as defined herein into contact with a mixture of one or more proteins.

The methods (i), (ii), (iii) and (vi) may also include the step of first providing the array according to one or more of the methods of the present invention.

Use of the proteins derived from the methods described herein form additional aspects of the invention. The skilled person will understand that a range of applications are known in the art in which modified proteins may be employed.

Thus, in further aspects, the present invention provides:

(i) the expression of tagged proteins produced by the methods of the invention in numerous expression hosts i.e. bacteria, yeast, mammalian cells (for example see, Walker EA, Clark AM, Hewison M, Ride JP, Stewart PM. Functional expression, characterization, and purification of the catalytic domain of human 11-beta -hydroxysteroid dehydrogenase type 1. J Biol Chem

2001 Jun 15;276(24):21343-50; Cai J, Daoud R, Georges E, Gros P. Functional Expression of Multidrug Resistance Protein 1 in Pichia pastoris. Biochemistry 2001 Jul 17;40(28):8307-16, and Hara H, Yoshimura H, Uchida S, Toyoda Y, Aoki M, Sakai Y, Morimoto S, Shiokawa K. Molecular cloning and functional expression analysis of a cDNA for human hepassocin, a liver-specific protein with hepatocyte mitogenic activity(1). Biochim Biophys Acta 2001 Jul 30;1520(1):45-53)

(ii) the use of a tagged protein produced by the methods as defined herein.

(iii) the use of a tagged protein produced by the methods as defined herein for analysis of interaction between expressed protein and other proteins within a yeast two-hybrid system via the cloning of said modified DNA molecule into a yeast two-hybrid expression vector (for example see, Staudinger J, Perry M, Elledge SJ, Olson EN. Interactions among vertebrate helix-loop-helix proteins in yeast using the two-hybrid system. J Biol Chem 1993 Mar 5;268(7):4608-11, and Vojtek AB, Hollenberg SM, Cooper JA. Mammalian Ras interacts directly with the serine/threonine kinase Raf. Cell 1993 Jul 16;74(1):205-14).

(iv) the use of a tagged protein produced by the methods as defined herein for immobilization on an affinity column/substrate, for example to allow the purification by affinity chromatography of, for example, a) interacting proteins, b) DNA or c) chemical compounds. (for example see, Rhodes N, Gilmer TM, Lansing TJ. Expression and purification of active recombinant atm protein from transiently transfected mammalian cells. Protein Expr Purif 2001 Aug;22(3):462-6; Zwicker N, Adelhelm K, Thiericke R, Grabley S, Hanel F. Strep-tag II for one-step affinity purification of active bHLHzip domain of human c-Myc. Biotechniques 1999 Aug;27(2):368-75, and Giuliani CD, Iemma MR, Bondioli AC, Souza DH, Ferreira LL, Amaral AC, Salvini TF, Selistre-de-Araujo HS. Expression of an active recombinant lysine 49

phospholipase A(2) myotoxin as a fusion protein in bacteria. Toxicon 2001 Oct;39(10):1595-600)

(v) the use of a tagged protein produced by the methods as defined herein in the immobilization by affinity purification for interrogation by antibodies (ELISA assay) as a diagnostic tool (for example see, Doellgast GJ, Triscott MX, Beard GA, Bottoms JD, Cheng T, Roh BH, Roman MG, Hall PA, Brown JE. Sensitive enzyme-linked immunosorbent assay for detection of Clostridium botulinum neurotoxins A, B, and E using signal amplification via enzyme-linked coagulation assay. J Clin Microbiol 1993 Sep;31(9):2402-9)

(vi) the use of a tagged protein produced by the methods as defined herein as a probe for a cDNA microarray to identify DNA binding proteins (for example, see DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996 Dec;14(4):457-60).

(vii) the use of a tagged protein produced by the methods as defined herein for elucidating the identity of proteins in the 'proteome' by mass spectrometric analysis of expressed protein components of source library or start material modified by the methods of the invention (for example, see Bordini E, Hamdan M. Investigation of some covalent and noncovalent complexes by matrix-assisted laser desorption/ionization time-of-flight and electrospray mass spectrometry. Rapid Commun Mass Spectrom 1999;13(12):1143-51).

Preferred features of each aspect of the invention are applicable to each other aspect, *mutatis mutandis*.

The present invention will now be described with reference to the following examples, which should not in any way be construed as limiting the scope of the invention.

FIGURE 1a: shows the construction of the vector pMM106H;

FIGURE 1b: shows details of the PCR amplification and exonuclease digestion of an example gene (GST) prior to tagging;

FIGURE 1c: shows details of the specific ligation and PCR amplification to introduce the tag;

FIGURE 1d: shows the reaction between Glutathione and 1-chloro-2,4-dinitrobenzene catalysed by GST.

Example 1

(a) Vector construction (see figure 1a)

The Inventors constructed a vector pMM106H derived from pUC19 which contains a strong hybrid promoter (P*trc*) to drive the expression of genes cloned into an *Nco* I site immediately downstream of the promoter sequence. The Inventors inserted a 676 bp nonsense DNA sequence as a stuffer fragment between the *Nco* I site and a downstream *Hpa* I site. *Hpa* I is a blunt-end cutter and is positioned to cleave the vector such that the downstream DNA encodes a polyasparagine, hexahistidine peptide if the reading frame is on the first base of the blunt-end. Following the hexahistidine tag is an amber stop codon (TAG) followed by the gene encoding the green fluorescent protein (GFP) of the jellyfish *Aequorea victoria*. The construction of pMM106H was confirmed by sequencing.

Genes cloned into pMM106H as *Nco* I/blunt-end fragments result in fusions to the His-tag and GFP only if the correct reading frame is created at the *Hpa* I site during cloning. GFP is used here as a reporter gene to facilitate visual screening of clones expressing the His tag, while also providing an indication of the correct folding of the fusion protein, since GFP is only active when folded into the correct conformation.

The amber stop codon will result in a small amount of the full length fusion protein for visualisation of green colonies, while most of the fusion protein will terminate immediately after the His tag and can be used for subsequent immobilisation and enzyme assays. It should be understood that a number of different peptides of proteins, whose soluble expression confers some observable phenotype on the cells, could be used in place of GFP as markers for expression and folding of the tagged proteins. These include, but are not limited to, chloramphenicol acetyl transferase, β-galactosidase, the lacZ fragment of β-galactosidase, and proteins capable of repressing transcription, such as the λ-CI repressor.

The template used in the procedure outlined below was pGSTN. This plasmid was constructed by first PCR-amplifying the *Schistosoma japonicum* glutathione *S* transferase (GST) gene from pGEX-2T (Pharmacia) under standard conditions using primers 'GSTfwd2' (5'-ATG CTG CAG ACG TCA ACA GTA TCC ATG GCC CCT ATA CTA GG-3') and 'GSTHindIII' (5'-GCG AGG AAG CTT GTC AAT CAG TCA CGA TGA ATT CCC G-3'). These primers introduce an *Nco* I restriction site at the start codon of GST, mutate the second residue of GST from serine to alanine, and introduce a stop codon in the multiple cloning site 3'- of the GST gene followed by a *Hin* dIII restriction site. The PCR product was then cloned under standard conditions as an *Nco* I/*Hin* dIII fragment into pTrcHisA (Invitrogen) previously digested with *Nco* I/*Hin* dIII to generate pGSTN.

(b) PCR amplification and exonuclease digestion of genes prior to tagging (see figure 1b)

The Inventors amplified the GST gene from the construct pGSTN using the polymerase chain reaction with custom-designed vector-specific primers 'STforward' (5'-ATG CTG ACG TCA TGA GGC CCA TGG GGC CCG GAT AAC AAT TTC ACA CAG G-3') and 'STreverse' (5'-GCG GAT CCT TGC GGC CGC CAG GCA AAT TCT GTT T-3') which bind to the vector 156 bp upstream of the start and 84 bp downstream of the stop codons respectively. 30 cycles of PCR (94°C 1min; 57°C 1min; 72°C 2min) were

carried out in four separate 100µl reactions. Each PCR reaction contained ~20ng template DNA, 50pmol each primer and 2.5 units *Pwo* polymerase. Each PCR reaction was carried out in a standard buffer (10mM Tris.HCl pH8.8, 25mM KCl, 5mM $(NH_4)_2SO_4$, 2mM $MgSO_4$, 10% DMSO). Each of the four PCR reactions then also contained a non-standard deoxynucleotide triphosphate mix, as follows:

Reaction 1) 200µM dATP, 200µM dTTP, 200µM dCTP, 150µM dGTP, 50µM α-S-dGTP;

Reaction 2) 200µM dATP, 200µM dTTP, 200µM dGTP, 150µM dCTP, 50µM α-S-dCTP;

Reaction 3) 200µM dATP, 200µM dGTP, 200µM dCTP, 150µM dTTP, 50µM α-S-dTTP;

Reaction 4) 200µM dGTP, 200µM dTTP, 200µM dCTP, 150µM dATP, 50µM α-S-dATP.

The amplification of the template DNA in the presence of α-S-dNTPs can of course be carried out by primer extension reactions using many different DNA polymerase, including thermostable polymerases which lack a 3' to 5' exonuclease activity, such as Tap polymerase, and non-thermostable polmerases, such as T4 DNA polymerase or the Klenow fragment of DNA polymerase I.

The inclusion of a single α-thio deoxynucleotide triphosphate in each specific PCR mix results in a random but statistical incorporation of the relevant α-S-dNTP into the specific final PCR product. These modified nucleotides are not substrates for Exonuclease III, and are used to halt the progressive removal of nucleotides by the enzyme. The four individual PCR mixes were then pooled, and purified using a QIAquick PCR cleanup kit (Qiagen), under standard conditions, and digested to completion with the restriction enzyme *Aat* II. The resulting ~1000bp PCR products were then gel-purified. Restriction with *Aat* II results in a 3'-overhang which is

resistant to Exonuclease III activity and, therefore, protects the 5' end of the PCR product from degradation.

Alternative methods for specifically protecting one end of the PCR product from exonuclease digestion can be readily envisaged including, but not restricted to the following. Any restriction enzyme which generates a 3'-overhang of 4 bases or more could be used in place of *Aat* II providing the requisite site is incorporated into the design of the PCR primers. Any restriction enzyme which generates a 5'-overhang could also be used in place of *Aat* II providing the requisite site is incorporated into the design of the PCR primers; in this case, generation of the 5'-overhang would be followed by a DNA-polymerase-mediated fill-in reaction in which the relevant α-thio dNTPs were used in place of the dNTPs such that the new 3'-end of the PCR product is now protected from exonuclease digestion.

10-15μg of the digested PCR product was then incubated with 75 units of Exonuclease III/μg of DNA for 30 minutes at 37°C in a 150μl reaction. The Exo III digestion was carried out in a standard reaction buffer (66mM Tris.HCl pH8.0, 6.6mM $MgCl_2$, 5mM DTT, 50μg/ml bovine serum albumin). These conditions ensure that digestion by Exo III has reached completion. The enzyme was then inactivated by heating to 75°C for 15 minutes. The product of the Exo III digestion is a nested set of deletions from the 3'-end of the PCR product.

Exonuclease III is a non-processive 3'- to 5'- exonuclease which is unable to hydrolyse α-thio-containing nucleotides so, in the present protocol, every time Exo III reaches an α-thio-deoxynucleotide base, the progressive truncation of the recessed 3'-end of the PCR product is halted. The net result is thus a nested set of deletions as a consequence of the random incorporation of each α-S-dNTP at the earlier stage. The ratio of α-S-dNTP to dNTP used in the original PCR amplifications was determined

empirically such that the envelope of nested deletions spanned a 400bp window of sizes centred approximately 100bp shorter than the original full length PCR product.

The size range of the truncations obtained can be controlled by altering the ratio of α-S-dNTP to normal dNTP. This is important when the method is applied to eukaryotic cDNAs because such cDNAs have variable length 3' untranslated regions, with the most common 3'-UTR length falling in the range of 200-300 bp. Since the relative efficiency of incorporation of each of the four α-S-dNTPs varies according to the identity of the polymerase, it is desirable to use α-S-dNTP to normal dNTP ratios which are optimised for each of the four bases and for the particular polymerase. Typically, the molar ratio of racemic α-S-dNTP to normal dNTP used will lie in the range 1:1 to 1:3.

(c) Removal of single-stranded regions and preparation for cloning (see figure 1c)

The nested set of deletions generated by Exonuclease III digestion in the previous step was purified by ethanol precipitation and resuspended in 1x mung bean nuclease buffer (50mM sodium acetate pH5.0, 30mM NaCl, 1mM $ZnSO_4$). The digested DNA was treated with (2units/μg) 30 Units of mung bean nuclease in a 100μl reaction at 30°C for 30 minutes. This process removed the 5'- and 3'-overhangs to yield blunt-end products. The reaction was stopped by the addition of EDTA to a final concentration of 5mM. The digested products were purified using a QIAquick PCR purification kit (Qiagen), digested with *Nco* I, and separated on a 1% agarose/TBE gel, using a 100bp DNA ladder as a standard. Products ranging in size from 800 to 1000bp were extracted from the agarose using a QIAquick gel extraction kit (Qiagen). Clearly, other single-strand nucleases such as S1 nuclease, can also be used to remove the 5'-overhang from the nested set of 3'-deletions generated by a 3'- 5'- exonuclease.

(d) Variation on the preparation of nested deletions

A number of other standard molecular biology methods for generating a nested set of deletions represent obvious variations on the original procedure. These include, but are not limited to, the use of any 3'- to 5'- exonuclease, any 5'- to 3'- exonuclease, or any endonuclease which truncates progressively from the termini of a linear DNA fragment. For example, the initial PCR amplification can be performed using the same reverse primer as above but with a forward primer which binds approximately 2 kb upstream of the start of the GST gene. This will generate a fragment in which the GST gene is flanked by >2 kb on the 5' end and only 84 bp on the 3' end. The purified PCR fragment can then be treated with Bal 31 nuclease, which progressively degrades linear duplex DNA from both the 5'- and 3'-ends. The enzyme is non-processive and the rate of degradation of the DNA depends on the time and temperature of the reaction, as well as the base composition of the DNA. Since the flanking region on the 3'-end of the GST gene in the PCR product is significantly shorter than that on the 5'end, degradation up to and beyond the stop codon will occur long before the start codon is reached from the other end. Time course experiments allow the optimum reaction conditions for removal of up to 400bp from the 3' end of the PCR product to be determined. The resulting nested set of deletions can then be blunt-ended to remove any remaining single-stranded regions, digested with a unique restriction enzyme encoded at the 5'-end of the gene by the original vector, and directionally cloned into the tag vector. Alternatively, Lambda exonuclease can be used to generate a nested set of 5'-deletions. The preferred substrate for this enzyme is 5' phosphorylated double stranded DNA so one end of the DNA substrate can be easily protected by having a 5' hydroxyl terminus.

The single-stranded 3' overhangs of the nested set of 5'- deletions generated by a 5'-3- exonuclease can be removed by a number of different enzymes, including T4 DNA polymerase or a single-stranded DNA specific nucleases such as RNAse T or Exonuclease T or mung bean nuclease.

(e) Cloning and analysis of the modified products (see figure 1c)

The vector pMM106H (3μg) was digested to completion with the restriction enzymes *Nco* I and *Hpa* I and the 2870bp backbone fragment was gel purified. The vector DNA and the restricted products prepared as described above were then ligated together under standard conditions and the ligation mix was used to transform *E. coli* DH5α cells which were then recovered and plated onto LB plates containing 100μg/ml carbenicillin.

This cloning procedure was carried out on the full set of deletions obtained in the previous step. However, only those deletions which excise the stop codon of the GST gene and end immediately after an in-frame codon should be able to give rise to in-frame fusions to the hexahistidine tag and GFP after cloning steps *via* this procedure; all other deletion products cloned in this manner should only lead to out-of-frame fusions to the hexahistidine tag and GFP or to unfused GST proteins, due to translational termination at the GST stop codon. This follows because ligation of the blunt end of the deletion product to the blunt end of the vector results in a genetic fusion in which the translation reading frame of the downstream vector DNA is dictated by the original reading frame of the GST coding region. If the deletion product ends in an incomplete codon, the newly appended hexahistidine-coding sequence will be out-of-frame with respect to the GST gene, whilst if the deletion product retains the GST stop codon, no translational fusion of GST to the hexahistidine tag will occur. The only hexahistidine-(and GFP-) tagged proteins which can arise from the overall process described above will therefore necessarily be GST fusions to the polyasparagine, hexahistidine tag. These will not necessarily be absolutely full-length clones, however their ability to fold correctly and retention of enzyme activity will be screened for in further steps.

Transformed colonies were visualised under UV light (365nm) and 30 colonies (approximately 10% of the total) which fluoresced green were selected by eye for further analysis. These colonies were replica-plated and analysed by colony Western blot under standard conditions using anti-His-tag and anti-GST antibodies. The anti-

His-tag antibody only binds to colonies which express a hexahistidine-tagged protein so the Western blot gives direct information about the number of colonies expressing in-frame fusions to the hexahistidine-tag. The anti-GST antibody, on the other hand, binds close to the C-terminus of the GST protein and therefore only recognises colonies expressing full- or nearly full-length GST proteins. The Inventors identified 19 colonies (63% of the green fluorescent colonies) containing protein which was positively recognised by both anti-His-tag and anti-GST antibodies. The DNA from 12 of these colonies was amplified, purified and sequenced. The sequencing data confirmed the presence of two perfect in-frame fusions to full length GST and 10 clones with short truncations in the GST gene, which were still in-frame with the hexahistidine tag. The frequency of isolation of full-length GST clones the Inventors obtained *via* this overall procedure is therefore approximately 17% (of the total number of green fluorescent colonies), while the frequency of isolation of full- or almost full-length GST clones, which are expected to retain activity, is approximately 63% (of the total number of green fluorescent colonies).

(f) Immobilisation and functional analysis of tagged proteins (see figure 1d)

E. coli DH5α cells were transformed with one of the full-length, hexahistidine-tagged GST plasmids created *via* the above methodology. A single carbenicillin-resistant colony was grown to mid-log phase in 10ml liquid culture and then supplemented with 100μM IPTG to induce expression of the hexahistidine-tagged GST. After growth for a further 4 hours, cells were harvested and lysed by freeze-thaw/lysozyme. SDS-PAGE of the crude lysate showed an overexpressed protein at the expected size (27kDa), which represented roughly 20% of total soluble protein, as well as a small amount of the 54 kDa GST-hexahistidine-GFP fusion, generated through amber suppression. The crude lysate (500μl; 100μg) was then mixed with Nickel-NTA magnetic beads (50μl; binding capacity 15μg hexahistidine-tagged protein) and the beads recovered by sedimentation under a magnetic field. The supernatant was discarded and the beads were washed and then resuspended in a glutathione *S*

transferase assay buffer containing 1mM each of glutathione and 1-chloro-2,4-dinitrobenzene. End point assay data was collected after 30 minutes at room temperature by measuring the absorbance at 340nm; this wavelength corresponds to the λ_{max} of the product of the GST-catalysed reaction.

As controls, cultures of DH5α containing either the parent vector (pMM106H) or a plasmid encoding an unrelated His-tagged protein (alanine racemase) were grown, induced, harvested, lysed and assayed in parallel. GST activity was only detected on the beads which had been mixed with the crude lysate containing the His-tagged GST, clearly demonstrating that the observed GST activity was due specifically to the immobilised His-tagged GST and moreover that the protein retained activity on specific immobilisation.

After completion of the enzymatic assay, protein was eluted from the magnetic beads by addition of buffer containing 100mM imidazole and analysed by SDS-PAGE. This showed that the sample which gave the positive activity assay result contained a single immobilised protein of the exact size expected for glutathione *S* transferase (27kDa), thus confirming that the observed activity on the beads was due to this recombinant His-tagged protein alone.

Example 2

(a) Vector construction

The Inventors constructed a second vector, pMM111, which is essentially the same as pMM106H (see Example 1), except that the 676bp *Nco* I/*Hpa* I nonsense DNA stuffer fragment is replaced with a 300bp *Nco* I/*Hpa* I fragment derived from the *Escherichia coli gdhA* gene; the *Hpa* I cloning site is replaced with a *Sma* I site, positioned such that the downstream hexahistidine tag is out of frame with the *gdhA* gene by 2

nucleotides; and the ATG start codon of the GFP gene is replaced with the codon for alanine (GCG). The vector has been designed such that an insert cloned into the *Sma* I site must contain the first nucleotide of a codon at its 3' end to put it in frame with the hexahistidine tag and GFP. The construction of pMM111 was confirmed by sequencing.

(b) Modification procedure to introduce tag

The Inventors then carried out a procedure identical to that described in Example 1 except for the following modifications. Firstly, only α-S-dTTP was incorporated in the original PCR amplification, i.e. reaction number 3 in section (b). Secondly, the final products were cloned in to the *Nco* I and *Sma* I sites of the vector pMM111.

This procedure has several theoretical advantages over that described in Example 1. These arise principally from the statistics associated with incorporation of a single α-thio dNTP in to the original PCR products. Thus, upon exhaustive exonuclease III digestion, the nested set of 3'-recessed deletions will all now terminate with a 3'-thymidine base rather than with any of the four nucleotides. Cloning of these fragments into the *Sma* I site of pMM111 will only result in an in-frame fusion to the hexahistidine tag and GFP if the 3'-T is either that of the first in-frame stop codon of the GST gene or precedes, and is in the same reading frame, as the 'T' of the first in-frame stop codon (this follows because digestion with *Sma* I results in a gap of 2 nucleotides before the coding sequence of the tag).

Statistically, 4-fold fewer nested deletions will be created by the exonuclease hydrolysis in this modified procedure than in Example 1. However, since all three possible stop codons contain 'T' as their first base, they will all be represented in the set of deletions and will therefore constitute a four-fold higher fraction of the full set of deletions. Given the probability of any given 'T' being in the same reading frame as the first 'T' of the stop codon, 33% of all clones resulting from this modified procedure should be in-frame fusions to the His tag encoded by the vector but those

deletions which affect folding (and hence function) give rise to 'white' colonies (for the reasons set out in Example 1). The Inventors have found that following this modified procedure, the fraction of precise, full length clones within the 'green' population is significantly higher than that found following the procedure in Example 1. The same argument hold of course for 'start' codons since all known start codons (ATG, GTG, TTG, ATT, CTG) contain a 'T' in the second position.

A further advantage of this modified procedure is that a polyA tail can be incorporated into the 5'-end of the forward primer used in the initial PCR amplification (e.g. Forward-A 5'-AAA AAA AAA AAA GAT CGA TCT C AT GAC GGA TAA CAA TTT CAC ACA GG-3'). During amplification with a dTTP:α-S-dTTP ratio of 3:1, there is a high probability then that at least one α-S-dTTP residue will be incorporated at the end of the complementary strand, at the 5' end of the PCR product. These incorporated nucleotides will be resistant to Exonuclease III digestion and will therefore remove the requirement for enzymatic steps in specifically protecting that end of the PCR product from degradation.

Example 3

(a) Modification of a second protein using the hexahistidine tag

Following the procedure as described in Example 1 for glutathione-*S*-transferase, the Inventors have demonstrated that the procedure is independent of the sequence of the gene being manipulated.

Thus starting with a plasmid encoding human transcription factor NF-κB p50 and following exactly the procedure described in Example 1 unless otherwise specified, the Inventors have been able to demonstrate the modification of NF-κB p50 such that the first in-frame stop codon has been excised and replaced by an in-frame fusion to DNA encoding a polyasparagine, hexahistidine tag and GFP (when the amber stop

codon is suppressed). The clones that fluoresced green, when excited with far uv light (365 nm) were further characterised. Colony Western blots using an anti-His-tag antibody allowed identification of clones expressing hexahistidine-tagged protein. The soluble protein lysates of these clones were resolved by SDS-polyacrylamide gel electrophoresis and probed with anti-His tag antibody. Immunoreactive signals were observed at approx. 65 kDa M_r (corresponding to translationally fused NF-κB p50 to the hexahistidine tag and GFP) and at approx. 38 kDa M_r (NF-κB p50-histag). In addition there was a signal at around 27 kDa M_r, which is probably a degradation product corresponding to the his tagged GFP protein. The sequencing data confirmed that several clones encoded perfect in-frame fusions of full- or almost full-length NF-κB to the hexahistidine tag. In a single experiment, 190 colonies were screened for green fluorescence. A total of 38 clones (20% of the total number of clones screened) fluoresced green when excited with far uv light (365 nm). Colony western blotting using anti-His tag antibody revealed that 29 of the 38 clones expressed the hexahistidine tag. Sequencing data confirmed that 18 of these clones were full-length, or close to full length, in-frame fusions of NF-κB p50 to the hexahistidine tag; 7 of these clones were absolutely full length, His tagged NF-κB p50 genes and the remaining 11 His-tagged clones had short truncations of 4 to 1 amino acid residues. This experiment clearly demonstrates the advantage of having a reporter system indicative of expression and proper folding of the protein. Roughly 50% of the clones that fluoresced green were full length in-frame fusions to a His tag, or had minor truncations which did not cross a domain boundary, fused in-frame to a His tag.

(b) Immobilisation and functional analysis of hexahistidine-tagged NF-κB p50

E. coli DH5α cells were transformed with one of the full-length, hexahistidine-tagged NF-κB plasmids created *via* the above methodology. A single carbenicillin-resistant colony was grown to mid-log phase in 10ml liquid culture and then supplemented with 100μM IPTG to induce expression of the hexahistidine-tagged NF-κB p50. After growth for a further 4 hours, cells were harvested and lysed by sonication. SDS-

PAGE of the crude lysate showed an overexpressed protein at the expected size (38kDa) which represented roughly 5% of total soluble protein.

κB motif 5'-CGT ATG TTG TGG GGA ATT CCC AGC GGA TAA C-3'
3'-GCA TAC AAC ACC CCT TAA GGG TCG CCT ATT G-5'
NF-κ B P50 binding site

A duplex oligonucleotide, 'κB motif', which contains a palindromic binding site for NF-κB p50, was labelled at the 3'-bases with digoxigenin using 3-terminal transferase under standard conditions.

The protein lysates were prepared using the lysozyme/freeze-thaw method in PBS (phosphate buffered saline pH 7.5) containing 5 mM β-mercaptoethanol. 200 µl of the soluble protein lysate from each clone, was applied to the Ni-NTA coated microwell and incubated at room temperature for 45 minutes. At the end of the incubation period, the wells were washed three times with PBST (PBS containing 0.02 % Triton X-100) to remove all the unbound proteins. The wells were washed three times with DNA binding buffer (10 mM Tris.HCl pH 7.4, 75 mM KCl containing 5 mM β-mercaptoethanol with a soak time of 1 minute. The 3' digoxigenin labelled κB motif (2 pmol) was added to the wells in 200 µl of the DNA binding buffer containing 1 µg of poly (dI-dC) non-specific DNA. After another 30 minutes incubation the unbound DNA was removed by washing the wells three times with 10 mM Tris.HCl pH 7.4, 25 mM KCl containing 0.02% Triton X-100. An anti-digoxigenin antibody-alkaline phosphatase conjugate was diluted to 150mU/ml in 'antibody dilution buffer' (10mM Tris.HCl pH7.4, 25mM potassium chloride) supplemented with 0.2% bovine serum albumin. The diluted antibody (200µl) was then applied to the microwells. After 30 minutes at room temperature, unbound antibody was removed by washing the microwells with 'antibody dilution buffer' (3x350µl) supplemented with 0.02% Triton X-100. 200µl of a buffer (100mM Tris.HCl pH9.5, 100mM NaCl, 50mM $MgCl_2$)

containing 250μM *p*-nitrophenyl phosphate (pNPP), an alkaline phosphatase substrate, was then added to the wells and the reaction allowed to proceed overnight at room temperature, after which the yellow colouration in each well (corresponding to formation of the product, *p*-nitrophenol) was quantitated at 405nm. The background rate of hydrolysis of the substrate pNPP was low so a positive assay result was therefore immediately clear from the appearance of yellow colour in the wells.

As controls in this assay the Inventors omitted either the crude lysate, or the labelled oligonucleotide, or the antibody, or added a 20-fold excess of unlabelled duplex oligo or replaced the hexahistidine-tagged NF-κB p50 containing crude lysates with equivalent amounts of a crude cell lysate from DH5α cells expressing hexahistidine-tagged GST in the same vector background.

In this assay, NF-κB p50 first binds to the labelled oligonucleotide *via* the specific binding site. The protein-DNA complex is then immobilised in the microwells *via* the hexahistidine tag and all other proteins (including complexes between the labelled oligo and other DNA binding proteins present in the crude lysate) together with any unbound, labelled oligo, are then washed away. Since the antibody-conjugate recognises the label on the oligo, not the hexahistidine-tagged protein, a positive signal in the assay can only be observed if the NF-κB p50-DNA interaction is maintained on immobilisation of NF-κB p50 *via* the tag; if this interaction is not maintained, the oligo will be lost during the washing steps so no colour change will be observed.

The Inventors found that the yellow product was only detected in the microwells which had contained the hexahistidine-tagged NF-κB p50 crude lysate and the digoxigenin-labelled oligonucleotide and to which the anti-digoxigenin antibody-alkaline phosphatase conjugate had been added. This demonstrated that the observed colour change was due specifically to the immobilised NF-κB p50-oligonucleotide complex and moreover that NF-κB p50 retained activity on specific immobilisation.

Example 4

(a) Identification of one protein from a pool of 10 genes

The Inventors have applied the procedure exactly as described in Example 1 except where specified to the pool of 10 different genes listed in the table below. The Inventors have generated arrays of the resultant specifically modified proteins such that each position in the array corresponds to a single recombinant protein immobilised through the tag appended as a result of this procedure. The Inventors have then screened the array by functional assay and have successfully identified individual protein components of the pool.

Table 1. Size and function of the ten genes in the pool

Gene	Size	Source and Function
glutathione *S* transferase	**950bp**	**bacterial; detoxification**
NF-κB p50	**1165bp**	**human; transcription factor**
maltose binding protein	**1325bp**	**bacterial; carbohydrate transport**
alanine racemase	**1342bp**	**bacterial; cell wall biosynthesis**
nuclear factor of activated T cells (NFAT)	**1087bp**	**murine; transcription factor**
indoleglycerolphosphate synthase	**1528bp**	**bacterial; amino acid biosynthesis**
phosphoribosylanthranilate isomerase	**920bp**	**bacterial; amino acid biosynthesis**
tryptophan synthase (α-subunit)	**1122bp**	**bacterial; amino acid biosynthesis**
chymotrypsin inhibitor 2	**389bp**	**barley; serine protease inhibitor**
β-lactamase	**1040bp**	**bacterial; antibiotic resistance**

Initially, all ten genes were subcloned in to the same pTrcHisA vector backbone since amongst other things this mimics the situation encountered with a cDNA library. The primers 'STforward' and 'STreverse' described in Example 1 were designed to be universal primers for the amplification of genes encoded within a pTrcHisA vector backbone.

The primer 'STforward' was designed such that it encodes a number of restriction sites as follows:

5'-ATG CTG ACG TCA TGA GGC CCA TGG GGC CCG GAT AAC AAT TTC ACA CAG G-3'

Aat II *Bsp* HI *Sfi* I

Thus, either of the restriction enzymes *Aat* II or *Sfi* I can be used to generate 3'-overhangs for exonuclease protection purposes. For directional cloning purposes at the end of the modification procedure, in this Example the Inventors chose to use *Bsp* HI since although statistically it will cut more frequently within a library, it generates cohesive ends which are compatible with the *Nco* I cloning site in the tag vector pMM106H used here and does not cut within any of the 11 genes in the present pool. Clearly, in principle any of the primer encoded restriction sites could be used providing that the tag vector contains an equivalent cloning site downstream of the promoter; *Sfi* I would have significant advantages in this regard in a larger library format because it has an 8bp recognition sequence so the frequency of random occurrence of an *Sfi* I site within a given gene will be much lower (1 in 6.5×10^4) than that for a 6bp recognition sequence such as that of *Bsp* HI (1 in 4,096).

The tag vector pMM106H is an 'ATG' vector, *i.e.* the 5'-cloning site (*Nco* I) overlaps the ATG start codon positioned downstream of a ribosome binding site (RBS) for expression of native proteins. However, in the procedure described here the Inventors are not reliant on the cloned genes having a common restriction site at their start codons. Instead, the Inventors simply rely on the vector-encoded promoter initiating transcription to produce mRNA, with the requisite signals for translational initiation being provided by the cloned genes themselves. Thus in this Example, all the genes in the original pool have a start codon immediately preceded by an RBS, irrespective of the presence or absence of a cloning site at the ATG. Since the primer 'STforward' binds upstream of the RBS in all eleven initial clones, subsequent post-modification cloning using any of the primer encoded restriction sites will introduce the newly

modified genes in to the tag vector together with their original RBS and ATG so translation initiation will be ensured. In a cDNA library format, a similar situation applies in that all full-length cDNAs will have their own 5'-untranslated regions (UTR) which contain the eukaryotic translational initiation signals. All that is required to obtain proper translational initiation then is to clone each modified cDNA together with its 5'-UTR in to a eukaryotic vector which provides transcriptional initiation signals so an equivalent universal set of PCR primers to those used in this Example could therefore be used in the modification of every member of a cDNA library in a single pot in a sequence-independent manner.

The experimental procedure was carried out as described in Example.1 with the following modifications. An equimolar pool of all ten genes was used as the template for initial PCR amplification using primers 'STforward' and 'STreverse', after which fragments were digested with *Aat II* to protect the 5' end. Exonuclease III and mung bean nuclease-treated fragments were generated exactly as in Example 1 and were then digested with *Bsp* HI, which restricts the fragments uniquely within the forward PCR primer binding site and generates a cohesive end for cloning into the vector pMM106H. The resulting fragments were gel purified and ligated in to the vector. Transformed cells were visualised under UV light (365nm) and colonies which fluoresced green were selected by eye for analysis by Western blot. Approximately 2% of the total number of transformed colonies fluoresced green. Of these, 103 (42%) expressed proteins which are recognised by anti-His tag antibodies. These colonies were inoculated individually into 1.5 ml of liquid medium in 96-deep-well blocks and grown overnight. Cells were harvested by centrifugation and lysed by freeze-thaw/lysozyme. The individual crude lysates were then applied to individual wells of a Nickel-NTA-coated 96-well plate and unbound proteins were removed by washing, leaving a discrete His tagged recombinant protein immobilised in each well. The immobilised proteins were then assayed for either GST or NF-κB activities using the assays described in Examples 1 and 3 and wells containing positive 'hits' were

identified in each case by the appearance of either green or yellow colouration respectively.

In the first assay, three proteins in the array were found to exhibit GST activity, giving a hit rate of approximately 3%. Sequencing of the corresponding plasmids revealed that all three encoded in-frame fusions between a GST gene and the hexahistidine tag and GFP gene; of these three, two were absolutely full-length GST and one was a slight truncation which clearly did not affect activity.

In the second assay, three proteins in the array showed positive 'κB-motif' DNA binding activity. Further characterisation of the positive clones, showed two of the clones were in-frame fusions of the NF-κB p50 gene to the hexahistidine tag, one of which was almost full length (truncated by one amino acid) whilst the other was more severely truncated but contained the entire DNA binding domain of NF-κB p50. Interestingly, since the assay is designed for binding to the cognate DNA sequence, truncations that still have the DNA binding domain intact, folded, and functional will be positive in the assay. The third clone was found to be an in-frame fusion of the DNA binding domain of the murine transcription factor NFAT to the His tag. The 3'-digoxigenin labelled 'κB motif' used in the assay contains a specific, high affinity (K_d approximately pM) binding site for NF-κB p50 but this same binding site is also specifically recognised by the DNA binding domain of NFAT with approximately nM affinity. This result therefore demonstrates that functional interrogation of arrays generated by this procedure can identify both specific interactions and also weaker interactions which are nonetheless specific and biologically relevant.

In a subsequent experiment, an array containing ca. 340 of His-tagged proteins was prepared according to the method of this Example. Analysis of the array by GST activity assay showed that 8% of all proteins in the array possessed strong GST activity. In addition, PCR analysis was carried out on a pool of the 340 encoding plasmid DNAs using gene-specific primers and this showed that each gene was represented in the His-tagged collection. These data therefore provide further

confirmation that the method of this example is sequence-independent and can be applied to a collection of different genes.

In summary therefore, the Inventors have used the procedures described in these Examples to create arrays of functional proteins in a microwell format and using these arrays the Inventors have successfully identified three different proteins from a pool based on either specific protein-ligand interactions (GST activity assay) or specific protein-DNA interactions (NF-κB binding assay).

CLAIMS

1. A method producing one or more proteins in which one or more domains are full length and correctly folded and which are each tagged at either the N- or C-terminus with one or more marker moieties, said method comprising:

(a) providing one or more DNA molecules having an open reading frame encoding said proteins together with 5' and/or 3' untranslated regions;

(b) amplifying said DNA molecules under conditions that statistically incorporate α-S-dNTPs as well as dNTPs into the daughter DNA molecules;

(c) specifically protecting the 5' or 3' end of said DNA molecules from nuclease digestion;

(d) treating said DNA molecules first with a 5' to 3'- or 3' to 5'-nuclease to generate a set of nested deletions followed by treating with a single-strand nuclease under conditions that allow removal of said 5' or 3' untranslated regions including the start or stop codons of said open reading frame;

(e) cloning the fragments generated by step (d) into an expression vector containing a coding sequence for one or more 5' or 3' marker moieties;

(f) expressing said encoded proteins.

2. A method as claimed in claim 1, wherein said amplification of said DNA molecule statistically incorporates a single α-S-dNTP.

3. A method as claimed in claim2, wherein the single α-S-dNTP is either α-S-dTTP or α-S-dATP.

4. A method as claimed in any one of claims 1 to 3, wherein said nuclease is exonuclease III or λ exonuclease.

5. A method as claimed in any one of claims 1 to 4, wherein said single-strand

nuclease is mung bean nuclease or T4 DNA polymerase.

6. A method as claimed in any one of claims 1 to 5, wherein the marker moiety allows confirmation of expression of said open reading frame.

7. A method as claimed in any one of claims 1 to 5, wherein the marker moiety allows confirmation of folding of said open reading frame.

8. A method a claimed in any one of claims 1 to 7, wherein the marker moiety encodes the green fluorescent protein.

9. A method as claimed in any one of claims 1 to 7 wherein the marker moiety is a peptide sequence, eg a hexa-histidine tag, a complete protein or protein domain, eg the maltose binding protein domain.

10 A method as claimed in claim 9 wherein the tag allows for purification of the individual proteins in the array.

11. A method as claimed in any one of claims 1 to 10 wherein the tag is inserted such that the start or stop codon for each of the proteins is replaced.

12. A method as claimed in any one of claims 1 to 10 wherein the tag is inserted in-frame in a region close to the terminus of each of the proteins which is unimportant for folding and function.

13. A method as claimed in any one of claims 1 to 10 wherein the tag is inserted in-frame within the open reading frame but in a region outside specific domain boundaries which is unimportant for folding and function.

14. A method as claimed in any one of claims 1 to 13 wherein amplification of said

DNA molecule in step (a) is by a non-proof-reading polymerase, e.g. Taq polymerase or the Klenow fragment of DNA polymerase I.

15. A method as claimed in any one of claims 1 to 14 wherein the ratio of said α-S-dNTPs to dNTPs is between 1:1 and 1:3

16. A method as claimed in any one of claims 1 to 15 wherein said 5' to 3' or 3'- to 5'- nuclease is unable to hydrolyse α-S-phosphodiester linkages.

17. A method as claimed in any one of claims 1 to 16 wherein said DNA molecule is a cDNA produced by reverse transcription from a mRNA sequence.

18. A method as claimed in any one of claims 1 to 17 wherein said method is carried out on multiple DNA molecules in parallel.

19. A method as claimed in any one of claims 1 to 18 wherein said method is carried out on a population of DNA molecules in a single pot.

20. A library of tagged proteins produced by the method of any one of claims 1 to 19.

21. A method for producing a protein array, said method comprising:
(a) clonally separating each member of the library of claim 20;
(b) expressing the individual tagged proteins in a spatially separated format;
(c) purifying each tagged protein by means of the marker moiety;
(d) depositing each protein in to a spatially defined array.

22. An array comprising proteins prepared by a method as defined in any one of claims 1 to 19 or produced by the method defined in claim 21.

23. An array as claimed in claim 22 wherein the components of the array are immobilised, eg to a solid surface.

24. An array as claimed in claim 22 or 23 wherein the individual proteins are immobilised by means of the tag moiety.

25. A method of screening one or more compounds for biological activity which comprises the step of bringing said one or more compounds into contact with a protein array as defined in any one of claims 22 to 24 and measuring binding of the one or more compounds to the proteins in the array.

26. A method of screening one or more proteins for specific protein-protein interactions which comprises the step of bringing said one or more proteins, eg a cell surface receptor, into contact with an array as defined in any one of claims 22 to 24, and measuring binding of the one or more specific proteins with the proteins of the array.

27. A method of screening one or more proteins for specific protein-nucleic acid interactions which comprises the step of bringing said one or more nucleic acid probes into contact with an array as defined in any one of claims 22 to 24, and measuring binding and measuring binding of the probes to the proteins in the array.

28. The use of an array as defined in any one of claims 22 to 24 in the rapid screening of a compound, protein or nucleic acid.

29. The use of an array as defined in any one of claims 22 to 24 in screening for molecules which recognise each protein in the array, wherein the molecules are preferably antibodies.

30. A method of generating an antibody array which comprises bringing a protein

array as defined in any one of claims 22 to 24 into contact with an antibody library, such that one or more proteins in the protein array bind to at least one antibody in the antibody library, removing any unbound antibodies and immobilisation of those antibodies bound to proteins in the protein array.

31. A method for the screening of protein function or abundance which comprises the step of bringing an antibody array as defined in claim 30 into contact with a mixture of one or more proteins.

32. A method as claimed in any one of claims 25 to 27 or 30 which also comprises the step of first providing the protein array as defined in any one of claims 22 to 24

33. The method as defined in claim 21, wherein the proteins in the array are purified and immobilised in a single step.

34. The use of a tagged protein produced by the methods as defined herein.

35. The use of a tagged protein produced by the methods as defined herein for analysis of interaction between expressed protein and other proteins

36. The use of a tagged protein produced by the methods as defined herein for immobilization on an affinity column/substrate, for example to allow the purification by affinity chromatography of, a) interacting proteins, b) DNA or c) chemical compounds.

37. The use of a tagged protein produced by the methods as defined herein in the immobilization by affinity purification for interrogation by antibodies (ELISA assay) as a diagnostic tool.

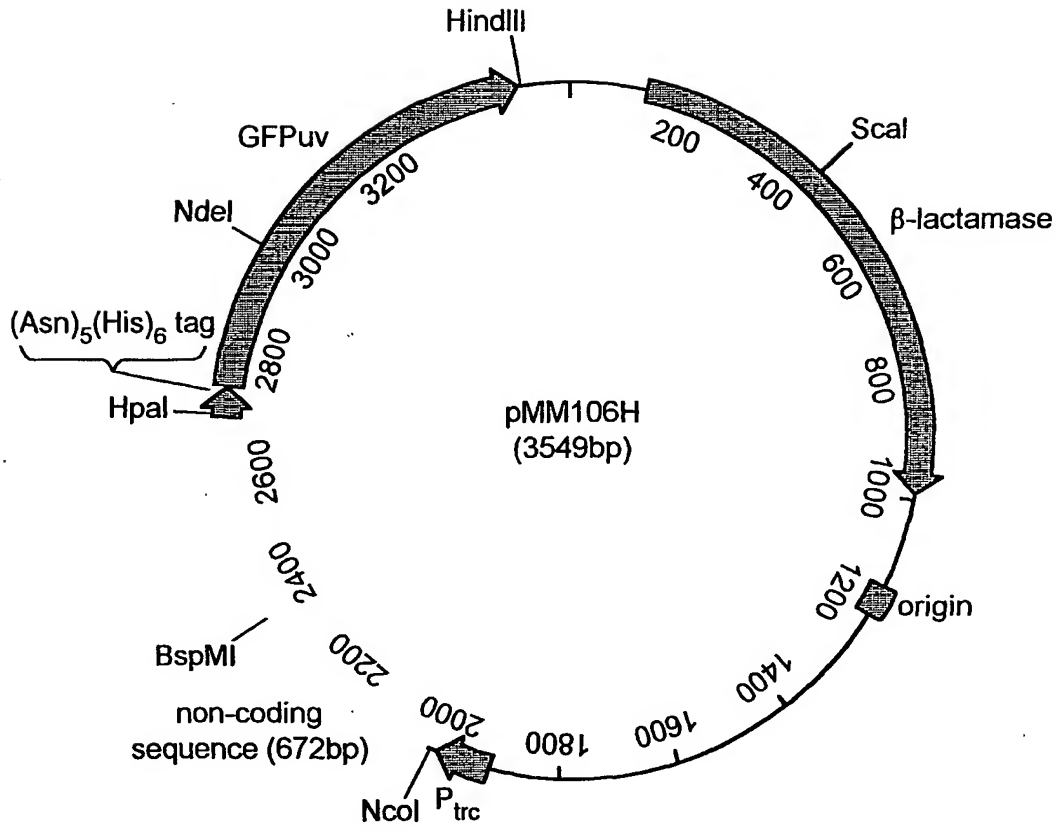
38. The use of a tagged protein produced by the methods as defined herein as a

probe for a cDNA microarray

39. The use as defined in claim 38 for the identification of DNA binding proteins.

40. The use of a tagged protein produced by the methods as defined herein for elucidating the identity of proteins in the 'proteome'

41. The use as defined in claim 40 wherein mass spectrometric analysis of expressed protein components of source library or start material modified by the methods of the invention is performed.




```
                                                    Amber
      Hpa I                                         Stop
_____GTTAACAACAACAACAACCACCACCACCACCACCACTAGGGCTCTATGAGT_____
      CAATTGTTGTTGTTGTTGGTGGTGGTGGTGGTGGTGATCCCGAGATACTCA
         N  N  N  N  N  H  H  H  H  H  H (E) G  S  M  S>
      _____ ASN-HIS TAG_____      ____GFP____
```

FIG. 1a

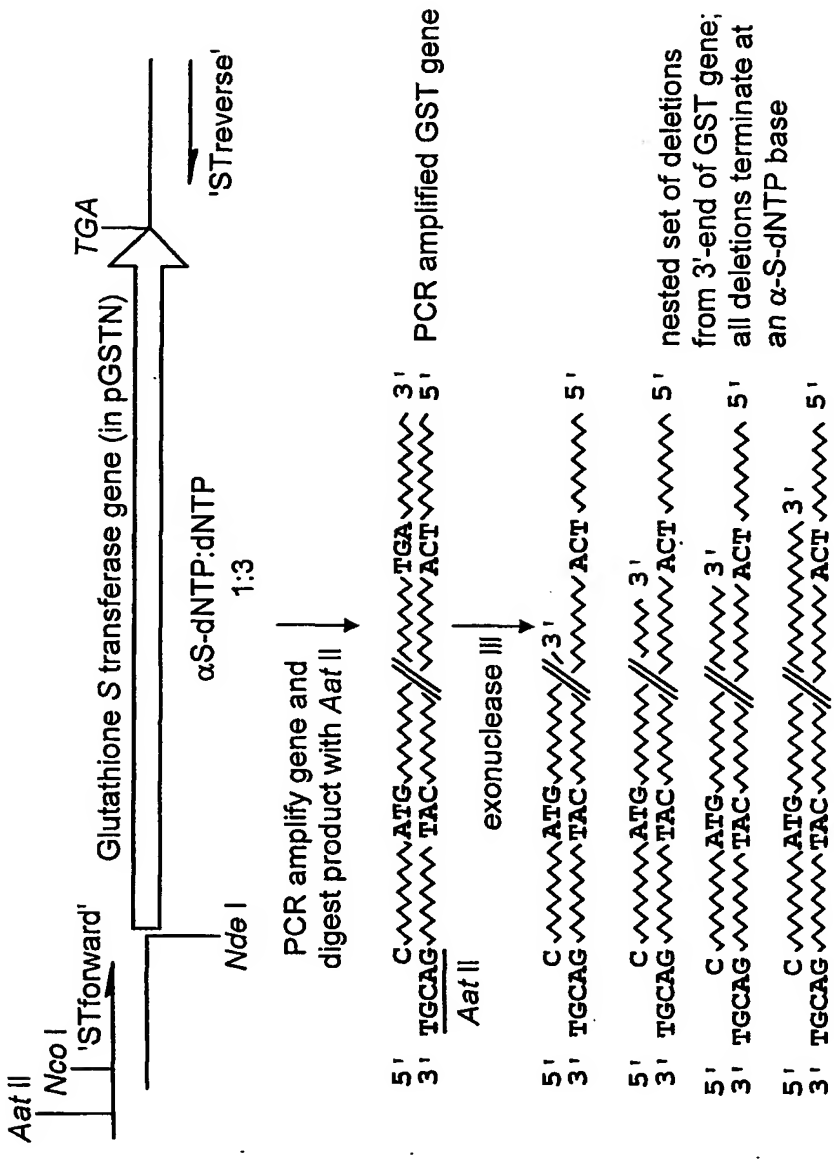

Aat II
Nco I
'STforward'
Glutathione S transferase gene (in pGSTN)
TGA
'STreverse'
Nde I
αS-dNTP:dNTP
1:3
PCR amplify gene and
digest product with Aat II
PCR amplified GST gene
5' C ATG 3'
3' TGCAG TAC 5'
Aat II
exonuclease III
nested set of deletions
from 3'-end of GST gene;
all deletions terminate at
an α-S-dNTP base

FIG. 1b

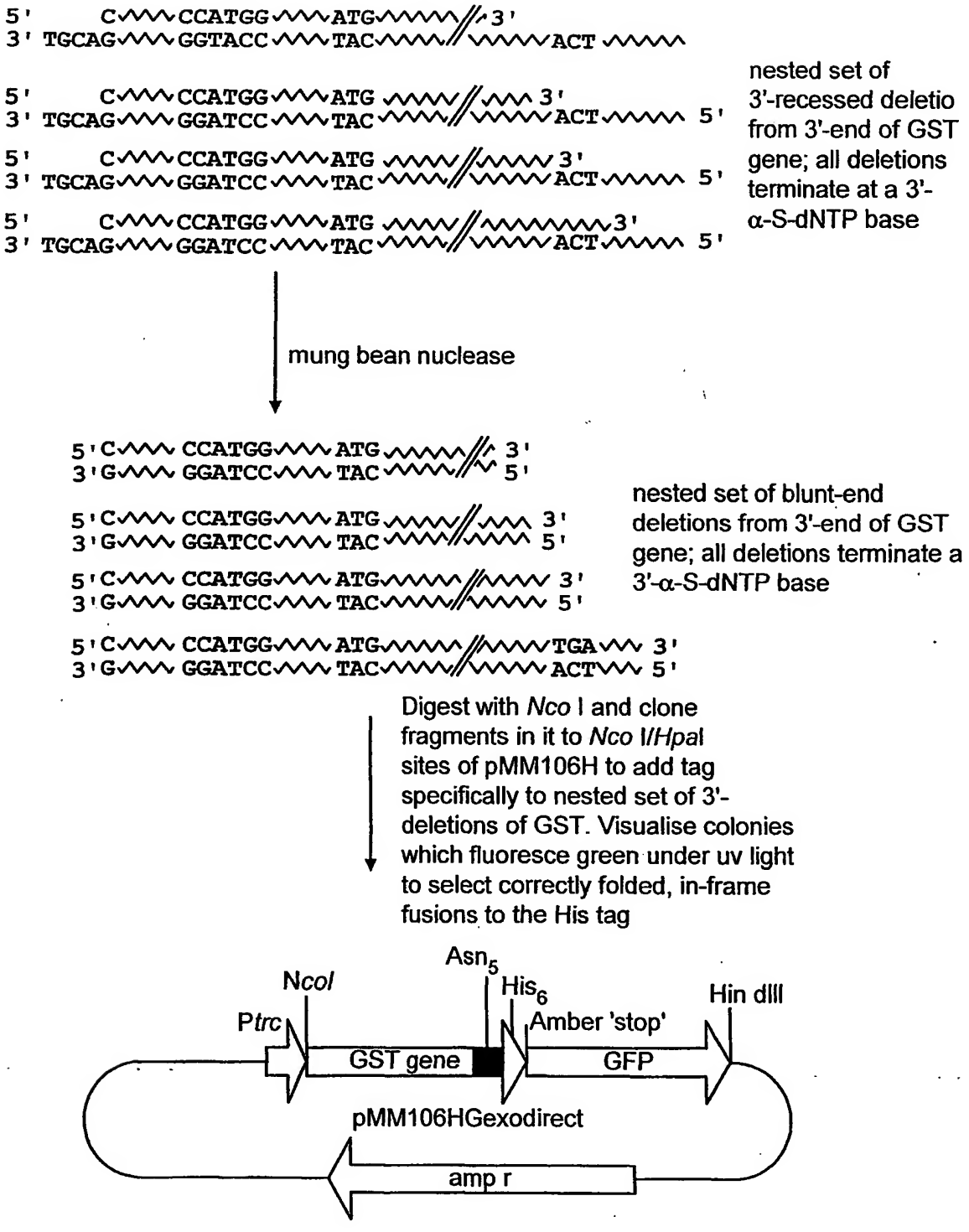

5' C CCATGG ATG 3'
3' TGCAG GGTACC TAC ACT
nested set of 3'-recessed deletio from 3'-end of GST gene; all deletions terminate at a 3'-α-S-dNTP base
mung bean nuclease
nested set of blunt-end deletions from 3'-end of GST gene; all deletions terminate a 3'-α-S-dNTP base
Digest with Nco I and clone fragments in it to Nco I/HpaI sites of pMM106H to add tag specifically to nested set of 3'-deletions of GST. Visualise colonies which fluoresce green under uv light to select correctly folded, in-frame fusions to the His tag
Asn5
NcoI
His6
Hin dIII
Ptrc
Amber 'stop'
GST gene
GFP
pMM106HGexodirect
amp r

FIG. 1C

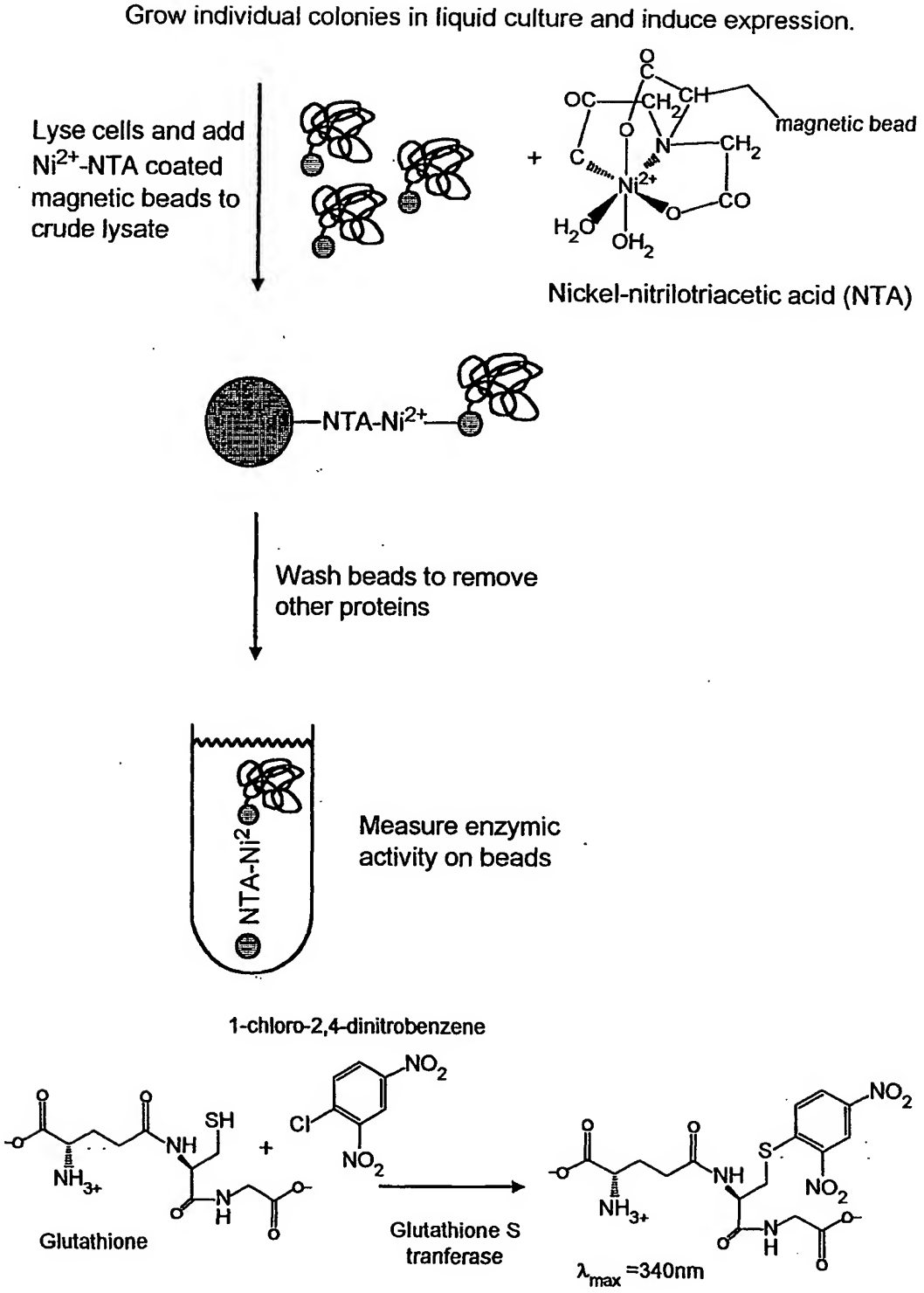

Grow individual colonies in liquid culture and induce expression.
Lyse cells and add Ni2+-NTA coated magnetic beads to crude lysate
magnetic bead
Nickel-nitrilotriacetic acid (NTA)
NTA-Ni2+
Wash beads to remove other proteins
NTA-Ni2
Measure enzymic activity on beads
1-chloro-2,4-dinitrobenzene
Glutathione
Glutathione S tranferase
λmax =340nm

FIG. 1d